DOCUMENT RESUME

ED 462 426 TM 033 690

AUTHOR	Zenisky, April L.; Hambleton, Ronald K.; Sireci, Stephen G.
TITLE	Effects of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics. MCAT Monograph.
SPONS AGENCY	Association of American Medical Colleges, Washington, DC.
REPORT NO	MCAT-5
PUB DATE	2001-12-00
NOTE	30p.
AVAILABLE FROM	Association of American Medical Colleges, Section for the Medical College Admission Test, 2450 N Street, NW, Washington, DC 20037. Tel: 202-828-0400; Fax: 202-828-1125; Web site: http://www.aamc.org/mcat.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Ability; College Applicants; College Entrance Examinations; Higher Education; Identification; *Item Response Theory; Test Items; *Validity
IDENTIFIERS	*Item Dependence; *Medical College Admission Test; Testlets

ABSTRACT

Measurement specialists routinely assume examinee responses to test items are independent of one another. However, previous research has shown that many contemporary tests contain item dependencies and not accounting for these dependencies leads to misleading estimates of item, test, and ability parameters. In this study, methods for detecting local item dependence (LID) are reviewed, and the use of testlets to account for LID in context-dependent item sets is discussed. LID detection methods and testlet-based item calibrations are applied to data from a large-scale, high stakes admissions test, and the results are evaluated with respect to test score reliability and examinee proficiency estimation. Data were from two forms of the Medical College Admission Test (MCAT) for 8,494 and 8,026 examinees. Results suggest the presence of LID impacts estimation of examinee proficiency. The practical effects of the presence of LID on passage-based tests are discussed, as are issues regarding the calibration of context-dependent item sets using item response theory. (Contains 3 figures, 11 tables, and 30 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

ED 462 426

Effects of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033690

MCAT Monograph 5

December 2001

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
P. Etienne
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

www.aamc.org/mcat

Effects of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics

April L. Zenisky, Ronald K. Hambleton, and Stephen G. Sireci

University of Massachusetts at Amherst

©Association of American Medical Colleges – Medical College Admission Test™
For reprint permission or information on the MCAT Monograph Series contact:
John H. Lockwood, Ph.D., Section for the MCAT, 2450 N Street NW, Washington DC 20037

Effects of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics

Abstract

Measurement specialists routinely assume examinee responses to test items are independent of one another. However, previous research has shown that many contemporary tests contain item dependencies, and not accounting for these dependencies leads to misleading estimates of item, test, and ability parameters. In this study, we (a) review methods for detecting local item dependence (LID), (b) discuss the use of testlets to account for LID in context-dependent item sets, (c) apply LID detection methods and testlet-based item calibrations to data from a large-scale, high stakes admissions test, and (d) evaluate the results with respect to test score reliability and examinee proficiency estimation. The results suggest the presence of LID impacts estimation of examinee proficiency. The practical effects of the presence of LID on passage-based tests are discussed, as are issues regarding how to calibrate context-dependent item sets using item response theory.

Introduction

The most basic unit of a test is the test item. Test development organizations spend more time and money developing and selecting items for inclusion on a test than on any other aspect of the test construction process. Numerous test items are needed to (a) adequately span the content or construct domain tested, and (b) provide reliable estimates of test takers' proficiencies. It has long been known that one way to increase test score reliability is to increase the number of items on a test. However, merely duplicating the same items will not accomplish the goal of reliable and valid measurement. Thus, test developers strive to develop items that provide unique information regarding test takers' knowledge, skills, and abilities. Redundancy among items is not desirable. Items that do not make a unique contribution to an assessment do not increase construct representation and exacerbate any construct-irrelevant factors that may be associated with an item, such as prior familiarity with the item context. For this reason, what is now known as local item dependence (LID) must be considered in the development and scoring of educational tests.

The concept of LID is best understood within the framework of item response theory (IRT). The most popular IRT models specify a single latent trait to account for all statistical dependencies among test items as well as all differences among test takers. It is this underlying trait, typically denoted theta (θ), that distinguishes items with respect to difficulty, and distinguishes test takers with respect to proficiency. The probability that a test taker will provide a specific response to an item is a function of the test taker's location on θ and one or more parameters (depending on the IRT model chosen) describing the relationship of the item to θ. Because IRT models are probabilistic, independence must be assumed, conditional on θ, between responses to any pair of items. This conditional independence is called local item independence

(Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968). When local item dependence is present on a test, inaccurate estimation of item parameters, test statistics, and examinee proficiency may result (Fennessy, 1995; Sireci, Thissen, & Wainer, 1991; Thissen Steinberg & Mooney, 1989). In addition, local item dependence introduces an additional (and generally unintended) dimension into the test at the expense of the construct of interest (Wainer & Thissen, 1996).

Several studies illustrated problems in not properly accounting for local item dependence. Thissen et al. (1989) and Sireci et al. (1991) analyzed items associated with reading passages and found that when the items were (improperly) treated as discrete, locally independent items, test information functions and reliability estimates were severely overestimated. This is an especially serious problem in computerized-adaptive testing (CAT), where the standard error of the estimate (SEE) is often used as the termination criterion. Since the SEE is the reciprocal of the test information, overestimating test information will result in premature termination of the test (Fennessy, 1995). Ferrara, Huynh, and Bagli (1997), Ferrara, Huynh, and Michaels (1999), and Yen (1993) investigated several potential causes of LID on performance assessments and found similar problems with respect to reliability estimation. In addition, these researchers provided several reasons for the existence of LID including multi-stage performance tasks, context-dependent item sets, and test speededness.

Classical test theorists were also concerned about inaccurate estimates that result when inter-item dependencies were not properly accounted for. For example, Kelley (1927), Guilford (1936), Thorndike (1951), Anastasi (1961), and others warned that items corresponding to a common stimulus or scenario (e.g., a set of items associated with a reading passage, table, figure, map, etc.) should all be placed into the same half-test when computing split-half reliability.

Otherwise, an inflated reliability estimate would occur, since these items were inter-dependent and the dependence would spuriously inflate the correlation between the two half-tests. Since coefficient alpha represents all possible split-halves, it would also be inflated by not accounting for such item inter-dependencies. Therefore, problems in not properly accounting for local item dependence are not limited to IRT.

Although local item dependence is undesirable, there are good reasons for including items that are inter-dependent on an assessment. Many real world tasks require solving related problems or solving a single problem in stepwise fashion. Thus, including context-dependent items on a test may increase construct validity. Examples of construct-relevant, inter-dependent items include items that require examinees to solve a problem and then explain how they arrived at their answer or the use of multiple items to measure comprehension of reading passages, scenarios, or graphs. Therefore, the challenge for the test developer is not the elimination of item dependencies, but rather how to properly model such dependencies so that local item dependence does not occur. Fortunately, several methods exist for detecting local item dependence, and for properly modeling construct-relevant local item dependencies within an IRT model.

In this paper, we apply different approaches to the detection of local item dependence on a large-scale, high stakes test: The Medical College Admissions Test (MCAT). Specifically, the purposes of this research are to investigate (a) the extent to which item dependencies exist in the multiple-choice test sections of the MCAT, (b) the impact of these item dependencies on reliability estimation, and (c) the use of testlet-based scoring in minimizing the negative consequences of these item dependencies. A study of the degree to which local item dependencies occur in multiple-choice data will permit the exploration of scoring methods that

may minimize such biasing effects. Seeing how serious this type of bias is with respect to item, test, and ability statistics when dichotomous scoring is used can provide useful information about the real psychometric quality of tests.

Modeling Testlet Structure To Ameliorate LID

If dependencies are found in the data when context-dependent item sets are used, one method by which those items could be scored is by the use of testlets and polytomous IRT models (Thissen, et al., 1989; Thissen, Billeaud, McLeod, & Nelson, 1997; Yen, 1993). A testlet is a scoring unit within a test that is smaller than a test, comprising items that may or may not be locally dependent (Wainer & Kiely, 1987). For example, a reading passage on the Verbal section of the SAT and its associated items could be construed as one testlet. A passage-based test could be composed of several such testlets. In using a polytomous IRT model to score testlets, the data can be analyzed while maintaining local independence across different testlets.

With respect to reliability estimation, the most accurate estimates are those in which items are locally independent, since item dependencies tend to inflate reliability estimation (Sireci et al., 1991). When seemingly distinct items related to a passage exhibit dependency, grouping them together into a testlet more properly models the test structure. Using this strategy, local item independence holds across testlets, since the testlet is modeled as a unit (i.e., a polytomous item). Thus, fitting sets of locally dependent items as testlets models the testlet-based structure of the test in a way that meets the local independence assumption of IRT.

One potential caveat to the use of polytomous IRT models could be a trade-off in information (Thissen, et al., 1997; Yen, 1993). By summing item scores within a testlet to compute testlet scores, information regarding the specific items examinees answered correctly is lost. In addition, fewer parameters are used to model the test compared to discrete-item scoring.

For example, if a 60-item test comprising ten, six-item testlets were scored dichotomously using the three-parameter IRT model, 180 item parameters would be estimated. In contrast, if the test were calibrated using a polytomous model to account for the testlet structure (e.g., Samejima's (1969) graded response model), only one discrimination parameter and 6 threshold parameters would be estimated for each testlet (a total of 70 parameters). Thus, some measurement information may be lost when collapsing items into testlets.

Given these tradeoffs in calibrating testlet-based tests, the best course of action may not be clear. The deciding factor is the extent to which dependencies in the data are consequential in terms of item, test, and ability statistics. When item dependencies are not present, forming the testlets and going to polytomous scoring does not improve anything. The potential benefits to be obtained in using testlets should be weighed against the added complexity in data analysis. Therefore, the degree to which LID exists on a test must be ascertained before deciding how to best model the test. Fortunately, effective methods for discovering LID exist.

LID Assessment Methods

Several different methods for assessing dependencies in dichotomous data have been developed. However, as Chen and Thissen (1997) pointed out, caution must be taken in the interpretation of the statistics provided by the methods as they exist for diagnostic purposes rather than hypothesis testing. Yen (1984) proposed the Q_3 statistic as an index of local item dependence. Q_3 is the correlation of the residuals for a pair of items after partialling out the trait estimate. To calculate Q_3, a proficiency estimate ($\hat{\theta}_a$) is calculated for each examinee and is used to estimate the expected performance of the examinee on each item (i.e., E_{ja}, where j denotes an item and a denotes an examinee). The residual (denoted d_{ja}) is calculated by taking

the deviation between an examinee's observed and expected performance on an item. Thus, for items j and j', Q_3 is the correlation of deviation scores across all examinees (i.e., $Q_{3jj'} = r_{(dj,dj')}$).

As examinee ability is used in both the calculation of the expected scores for examinees (in E_{ja} by way of $\hat{\theta}$) and also observed scores, this duplication (termed part-whole contamination by Kingston & Dorans, 1982), tends to produce Q_3 values that are marginally negative. When no local dependence exists, the expected value of Q_3 is $-1/(n-1)$, where n is the number of items on the test. In practice, this statistic has been used successfully by Yen (1993), Fennessy (1995), and Chen and Thissen (1997).

Another index suggested for identifying LID in practice is the directionally signed G^2 statistic, distributed normally as χ^2 with 1 degree of freedom (Bishop, Fienberg, & Holland, 1975; Chen & Thissen, 1997). The G^2 statistic is the likelihood ratio test:

$$G^2 = -2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij} \ln\left(\frac{E_{ij}}{O_{ij}}\right) \quad . \tag{4}$$

This LID statistic has been compared to Yen's Q_3; while both could detect dependencies with some power, Q_3 seemed to outperform G^2 for the most part (Chen & Thissen, 1997).

Conditional inter-item correlations have also been proposed as a measure of LID (Ferrara, Huynh, & Baghi, 1997; Ferrara, Huynh, & Michaels,1999; Huynh & Ferrara, 1994). In this method, examinees are sorted into (typically eight to ten) groups based on total test score, and inter-item correlations are computed within each test score interval. The inter-item correlations within a testlet can be averaged across each score level and each item to obtain a statistical measure of LID for each testlet. This measure of within-testlet LID can be compared to the same statistic computed across testlets. If the average within-testlet correlations are higher than the between-testlet correlations, reliability estimates derived from dichotomous scoring of

the items will be positively biased. Lee and Frisbie (1999) also computed average within- and between-testlet correlations in their generalizability theory approach to assessing the reliability of tests composed of testlets. When testlet scoring was used on the sets of items in their research, the difference between the computed passage reliability and the generalizability coefficient was small, supporting the position that testlet scoring was the appropriate level of scoring to use, as compared to dichotomous item scoring.

Wainer and his colleagues (Sireci et al., 1991; Wainer, 1995; Wainer & Thissen, 1996) also demonstrated that the presence of LID on a test can be ascertained by comparing two separate reliability estimates. The first estimate assumes all items are locally independent and ignores the testlet structure. The second estimate models the inherent testlet structure, which involves forming testlets for all context-dependent item sets. If the testlet-based reliability estimate is substantially lower than the item-based estimate, LID is present.

In this paper, we employ two methods for detecting LID. First, we model context-dependent item sets using testlets and compare the resulting reliability estimates to those obtained when the test is considered to only comprise locally independent items. Second, we calculate Q_3 statistics among the items. Our analyses span two test forms and three different content sections of the MCAT.

Method

Data

Data from a 1994 administration of the MCAT were used in these analyses. Examinee responses for each of three multiple-choice test sections (Verbal Reasoning, Biological Sciences, and Physical Sciences) were analyzed. There were two forms for each test section, differing only in the ordering of item sets. These different orderings of items sets were used to discourage

examinees from copying each other's answer sheets. The ordering of items within the item sets did not change between the two forms. In comparing the two orderings it can be determined if the ordering of the item sets impacts within-passage local dependence. On Form 1, data for 8,494 examinees were available, and on Form 2 there were 8,026 examinees. On both forms of this MCAT, the Verbal Reasoning test section comprised eight passages (55 total items). The Biological Sciences and Physical Sciences test sections both had nine passages and eleven discrete items (63 total items). All passages were followed by a set of items directly relating to the passage. Examinee responses for each item were scored either right or wrong. Omitted or not reached items were scored as wrong, which was consistent with the operational scoring of the test.

Data Analyses

Reliability Analyses

Coefficient α and IRT marginal reliability estimates (Green, Bock, Humphreys, Linn, & Reckase, 1984) were computed for data scored dichotomously as well as data scored polytomously. Two strategies were used to compute these reliability estimates for each test section. The first strategy was based on "traditional" scoring where all items were treated as discrete and were scored dichotomously. The other estimate was based on scoring the testlets polytomously. In this testlet-based scoring, an examinee's score on a testlet was computed by adding up the number of items within the testlet s/he answered correctly. Comparing the reliability estimates provided by these two scoring schemes provides a measure of the degree of LID due to items measuring a common passage. For example, if the testlet-based reliability coefficient is lower than the coefficient based on dichotomous scoring, the latter coefficient is probably an overestimate (Sireci, et al., 1991, Thissen, et al., 1989). However, as Sireci, et al.

(1991) pointed out, some drop in reliability is expected due to the fact that there are fewer "items" when testlets are formed from discrete items. Therefore, for the purpose of comparison, testlets were also formed randomly (i.e., joining items together from different passages) to gauge the drop in reliability due to the process of forming testlets. This "faking" of testlets was originally used by Yen (1993) for this same purpose.

Q_3 Analyses

The dichotomously scored data were calibrated using the three-parameter logistic IRT model. Yen's (1984) Q_3 statistics were used to assess dependencies within "true" testlets (i.e., passage-based testlets) and "fake" testlets (i.e., testlets formed randomly) for each test section and test form. The Q_3 statistics computed from the fake testlets provided a baseline for evaluating the magnitude of LID found in the other analyses, as proximate items randomly grouped together should not exhibit any LID. The Q_3 matrix for each test section was computed using the IRTNEW program (Chen, 1998). Summary statistics were then compared. The Q_3 values and the summary statistics were inspected for patterns relating to ordering effects, item sets, and passage types.

Ability Estimation

Plots of ability estimates from dichotomous and polytomous scoring allowed for a study of the impact of scoring method on the calculation of ability scores. Data sets where testlet scoring was used were calibrated using MULTILOG (Thissen, 1991). The choice of polytomous IRT model was not difficult because researchers have found that the two commonly-used polytomous IRT models, Samejima's (1969) graded response model and the generalized partial credit model (Muraki, 1992), provide highly similar results when used to analyze data with

responses in multiple categories (Maydeu-Olivares, Drasgow, & Mead, 1994; Tang & Eignor, 1997; Thissen, et al, 1997). In this study, the graded response model was used.

Results

Reliability Analyses

A summary of the coefficient α reliability analyses is presented in Table 1. Three sets of estimates are provided for each form of each test section: traditional α reliability estimates based on scoring all of the items dichotomously, "true" testlet-based reliabilities calculated using testlet scores for all passage-based (i.e., context-dependent) items, and "fake" testlet-based reliabilities calculated by summing together items that were randomly grouped to form testlets. It should be noted that the Biological Sciences and Physical Sciences sections included nine testlets and 11 dichotomously scored items, whereas the Verbal Reasoning section comprised eight testlets.

Table 1. Coefficient α Reliabilities

Test Section/ Form	Dichotomous Items		"True" Passage-Based Testlets	"Fake" (Randomly Formed) Testlets	# Items in Testlet Scoring
	# Items	α			
Ver.Reas. 1	55	.85	.79	.85	8
Ver.Reas. 2	55	.87	.82	.87	8
Bio.Sci. 1	63	.86	.82	.83	20*
Bio.Sci. 2	63	.87	.83	.84	20*
Phys.Sci. 1	63	.87	.83	.85	20*
Phys.Sci. 2	63	.89	.84	.84	20*

*Nine testlets and 11 discrete items

For most test sections and forms, no differences were observed between reliability estimates computed from the dichotomously scored data and those computed from the "fake" testlets. In contrast, the estimates for the dichotomously scored data tended to be larger than those for the context-dependent testlets. These results indicate some LID in the data. The Spearman-Brown formula is one way in which reliability estimates for tests can be compared to determine the size of the overestimate of reliability in the dichotomous case (Sireci, et al., 1991; Wainer, 1995). This measure provides an estimate of the amount by which a testlet-based test

would need to be lengthened to obtain the same reliability as the dichotomously scored test. Table 2 highlights the bias in the original dichotomously scored test reliability estimates.

Table 2. Spearman-Brown Length Increase Statistics
(from Coefficient α Reliability Estimates)

Test Section/Form	Length Increase ("True" Testlets)	Length Increase ("Fake" Testlets)
Ver.Reas. 1	1.51	1.00
Ver.Reas. 2	1.47	1.00
Bio.Sci. 1	1.35	1.26
Bio.Sci. 2	1.37	1.27
Phys.Sci. 1	1.37	1.18
Phys.Sci. 2	1.55	1.54

The results in Table 2 suggest that the reliability estimates based on dichotomous scoring of the Verbal Reasoning passages are inflated due to LID. For Verbal Reasoning, a 50% increase in the testlet-based test would be needed to achieve the level of reliability (falsely) indicated by the dichotomous analysis. For the other test sections, the length increase for the true testlets was similar to the length increase for the fake testlets, which suggests the drop in reliability may be due to the process of forming the testlets as opposed to LID.

In addition to coefficient α reliability estimates, IRT-based marginal reliability estimates were computed by applying the three-parameter logistic model to the dichotomously scored items, and the graded response model to the polytomously-scored testlets. These marginal reliability estimates are reported in Table 3. The test length increases needed to achieve the level of reliability estimated from the dichotomous data are presented in Table 4. The results tell essentially the same story as the α reliabilities. LID dependence appears to be most prevalent on the Verbal Reasoning section. Due to weighting of item scores within IRT, the marginal reliabilities have a tendency to be slightly higher than coefficient α, but generally within 0.02 (Wainer & Thissen, 1996).

Table 3. IRT Marginal Reliabilities

Test Section/ Form	Dichotomous Items		"True" Testlets	"Fake" Testlets	# Items in Testlet Scoring
	# Items	α			
Ver.Reas. 1	55	.87	.81	.85	8
Ver.Reas. 2	55	.88	.83	.87	8
Bio.Sci. 1	63	.88	.85	.86	20
Bio.Sci. 2	63	.89	.86	.87	20
Phys.Sci. 1	63	.89	.86	.87	20
Phys.Sci. 2	63	.90	.87	.87	20

*Nine testlets and 11 discrete items

Table 4. Spearman-Brown Length Increase Statistics (from IRT Marginal Estimates)

Test Section/Form	Length Increase ("True" Testlets)	Length Increase ("Fake" Testlets)
Ver.Reas. 1	1.57	1.18
Ver.Reas. 2	1.77	1.10
Bio.Sci. 1	1.29	1.19
Bio.Sci. 2	1.32	1.21
Phys.Sci. 1	1.32	1.21
Phys.Sci. 2	1.35	1.35

Local Dependence Assessment

Test-Section Level Q_3 Analyses

The Q_3 matrix was obtained for each of the test sections and forms. Using these matrices, the mean Q_3 value for each test section and form was computed by averaging Q_3 values for pairs of items located within the same testlet. Table 5 presents these means for both the real and fake testlets. In addition, the expected value of the Q_3 for each test section, which assumes the items are locally independent, is also presented.

Table 5. Mean Q_3 Statistics for Test Sections

Test Section	"Fake" Testlets		"True" Testlets	
	Form 1	Form 2	Form 1	Form 2
Verbal Reasoning (expected Q_3: -.019)	-.026	-.018	.024	.032
Biological Sciences (expected Q_3: -.016)	-.019	-.015	.010	.013
Physical Sciences (expected Q_3: -.016)	-.013	-.020	.018	.013

BEST COPY AVAILABLE

The mean Q_3 values for the fake testlets closely approximated the expected values, while the mean Q_3 values observed for the true testlets are elevated. Consistent with the reliability differences noted earlier, the greatest disparity between observed and expected results occurs within the Verbal Reasoning test section, with lesser differences exhibited by Biological and Physical Sciences.

Testlet-Level Q_3 Analyses

Next, mean Q_3 values for items within each testlet for each test section and test form were computed. These statistics, ranged from -.004 to .058 for the true testlets and from -.030 to -.009 for fake testlets. For the true testlets, the mean Q_3 values always exceeded the expected value. In contrast, the mean Q_3 values for the fake testlets closely approximated the expected Q_3 values. These computations provide definitive evidence for the presence of statistical LID, though actual levels vary across test sections and item sets.

Tables 6 and 7 provide the mean Q_3 values for item sets on the Verbal Reasoning section. As expected, the mean Q_3 statistics for the fake testlets were negative, correctly indicating the absence of LID. In comparison, the mean Q_3 statistics for the true testlets are positive (ranging from .009 to .058 across the two forms). While the magnitude of the dependence varies across the different testlets, these results clearly suggest passage-based LID exists within this test section. Of particular note is the somewhat higher level of dependence observed for the last testlet administered on each form. A mean Q_3 statistic that is higher for a testlet administered near the end of the test is suggestive of dependence due to speededness, as items near the end of a test can exhibit LID by virtue of their positioning. On the Verbal Reasoning section, the positioning of testlets 7 and 8 were interchanged across the two forms. The mean Q_3 value for the last testlet on Form 1 was .052. The mean Q_3 value for this testlet on Form 2, when it was

the second-to-last passage, was .043. Similarly, the mean Q_3 value for the last testlet on Form 2 was .042 on that form, but .021 when it was the second-to-last testlet on Form 1. These results suggest that some, but not all, of the LID noticed in these testlets may be due to speededness.

Table 6. Mean Q_3 Statistics for Testlets and Deviation from Expected Verbal Reasoning Form 1 (Expected Q_3: -.019)

	No. of Items	"True" Testlets (Deviation)	"Fake" Testlets (Deviation)
Testlet 1	10	.015 (.034)	-.015 (.004)
Testlet 2	6	.032 (.051)	-.030 (-.011)
Testlet 3	7	.010 (.029)	-.016 (.003)
Testlet 4	6	.031 (.050)	-.028 (-.009)
Testlet 5	6	.008 (.027)	-.048 (-.029)
Testlet 6	6	.022 (.041)	-.022 (-.003)
Testlet 7	6	.021 (.040)	-.030 (-.011)
Testlet 8	8	.052 (.071)	-.018 (.001)
	Mean	.024 (.043)	-.026 (-.007)

Table 7. Mean Q_3 Statistics for Testlets and Deviation from Expected Verbal Reasoning Form 2 (Expected Q_3: -.019)

	No. of Items	"True" Testlets (Deviation)	"Fake" Testlets (Deviation)	Form 1 Order
Testlet 1	6	.030 (.049)	-.016 (.003)	4
Testlet 2	7	.030 (.049)	-.018 (.001)	3
Testlet 3	6	.026 (.045)	-.007 (.012)	6
Testlet 4	6	.009 (.028)	-.016 (.003)	5
Testlet 5	10	.014 (.033)	-.017 (.002)	1
Testlet 6	6	.058 (.077)	-.028 (-.009)	2
Testlet 7	8	.043 (.062)	-.029 (-.010)	8
Testlet 8	6	.042 (.061)	-.013 (.006)	7
	Mean	.032 (.051)	-.018 (.001)	

Tables 8 and 9 presents the mean Q_3 values for Forms 1 and 2 of the Biological Sciences section. A few testlets appear to contain some LID. Testlet 8 on Form 2 exhibited the largest mean Q_3 value (.044). Its counterpart on Form 1, testlet 5, also exhibited the largest Q_3 (.043). Unlike the Verbal Reasoning section, these relatively larger Q_3 values were not consistent with a speededness hypothesis.

BEST COPY AVAILABLE

Table 8. Mean Q_3 Statistics for Testlets and Deviation from Expected Biological Sciences Form 1 (Expected Q_3: -.016)

	No. of Items	"True" Testlets (Deviation)	"Fake" Testlets (Deviation)
Testlet 1	6	-.004 (.012)	-.012 (.004)
Testlet 2	6	.007 (.023)	-.022 (-.006)
Testlet 3	5	.001 (.017)	-.030 (-.014)
Testlet 4	5	-.003 (.013)	-.026 (-.010)
Testlet 5	7	.043 (.059)	-.013 (.003)
Testlet 6	5	.023 (.039)	-.012 (.004)
Testlet 7	7	.012 (.028)	-.033 (-.017)
Testlet 8	5	.017 (.033)	-.006 (.010)
Testlet 9	6	.004 (.020)	-.013 (.003)
Mean (Deviation)		.010 (.027)	-.019 (-.003)

Table 9. Mean Q_3 Statistics for Testlets and Deviation from Expected Biological Sciences Form 2 (Expected Q_3: -.016)

	No. of Items	"True" Testlets (Deviation)	"Fake" Testlets (Deviation)	Form 1 Order
Testlet 1	5	.005 (.021)	-.008 (.008)	3
Testlet 2	5	-.004 (.012)	-.018 (-.002)	4
Testlet 3	6	.007 (.023)	-.022 (-.006)	2
Testlet 4	6	.006 (.022)	-.008 (.008)	1
Testlet 5	7	.009 (.025)	-.007 (.009)	7
Testlet 6	5	.021 (.037)	-.024 (-.008)	8
Testlet 7	5	.020 (.036)	-.020 (-.004)	6
Testlet 8	7	.044 (.060)	-.019 (-.003)	5
Testlet 9	6	.008 (.024)	-.007 (.009)	9
Mean		.013 (.029)	-.015 (.001)	

The mean Q_3 values for the last two testlets on both forms of the Physical Sciences section were slightly elevated, which initially suggested speededness. However, the mean Q_3 values for these testlets remained relatively large when these passages were placed earlier in the test. For example, the same passage had the largest mean Q_3 value on both forms. This value was .080 on Form 2 when the passage was the eighth (second-to-last) testlet, and .048 when it was the fifth testlet on Form 1. Thus, part of the LID noted on Form 2 may be due to speededness, but clearly context-dependence may also be a cause of LID within this testlet. The

BEST COPY AVAILABLE

mean Q_3 value for the fake Physical Sciences testlets were close to their expected values on both forms. Thus, the LID observed in this test section may be related to both passage and speededness effects. Tables 10 and 11 present the mean Q_3 values for item sets on Forms 1 and 2 of the Physical Sciences section.

Table 10. Mean Q_3 Statistics for Testlets and Deviation from Expected Physical Sciences Form 1 (Expected Q_3: -.016)

	No. of Items	"True" Testlets (Deviation)	"Fake" Testlets (Deviation)
Testlet 1	6	.001 (.017)	-.024 (-.008)
Testlet 2	7	.009 (.025)	-.009 (.007)
Testlet 3	5	-.004 (.012)	-.018 (-.002)
Testlet 4	5	.004 (.020)	-.009 (.007)
Testlet 5	6	.048 (.064)	-.013 (.003)
Testlet 6	6	-.001 (.015)	.002 (.018)
Testlet 7	6	.020 (.036)	-.021 (-.005)
Testlet 8	6	.045 (.061)	-.019 (-.003)
Testlet 9	5	.042 (.058)	-.010 (.006)
	Mean	.018 (.034)	-.013 (.003)

Table 11. Mean Q_3 Statistics for Testlets and Deviation from Expected Physical Sciences Form 2 (Expected Q_3: -.016)

	No. of Items	"True" Testlets (Deviation)	"Fake" Testlets (Deviation)	Form 1 Order
Testlet 1	5	-.001 (.015)	-.021 (-.005)	4
Testlet 2	5	.006 (.022)	-.026 (-.010)	3
Testlet 3	7	.017 (.033)	-.022 (-.006)	1
Testlet 4	6	-.006 (.010)	-.012 (.004)	2
Testlet 5	6	.008 (.024)	-.012 (.004)	7
Testlet 6	6	-.028 (-.012)	-.026 (-.010)	8
Testlet 7	6	.003 (.019)	-.015 (.001)	6
Testlet 8	6	.080 (.096)	-.017 (-.001)	5
Testlet 9	5	.040 (.056)	-.027 (-.011)	9
	Mean	.013 (.029)	-.020 (-.004)	

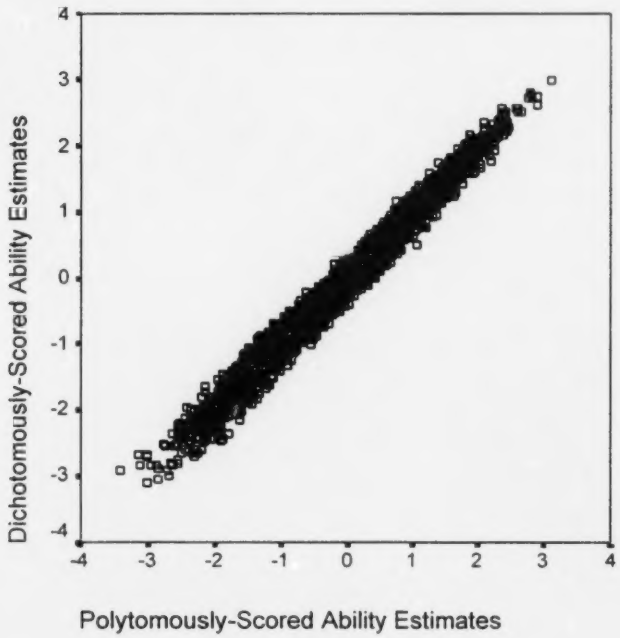
Ability Estimation

The ultimate purpose of this assessment of LID and experimentation with polytomous IRT models is to make informed decisions about the true impact of LID on ability estimation. When passage-related dependence is observed on a test or test section, the degree to which it

BEST COPY AVAILABLE

distorts ability estimates must be discovered and interpreted. Should it be severe enough to preclude valid usage of examinee test scores in the manner in which they were intended, then the use of testlet scoring is warranted. Of course, "severe enough" requires a judgment, but it is important to recall that the process of interpreting dependence itself is a somewhat imprecise exercise. LID analyses are largely exploratory in nature, and are completed to provide guidance for the test developer.

Figure 1. Plots of Ability Estimates: Dichotomous and Polytomous Scoring
Biological Sciences Form 1 (Correlation: 0.990)



Figures 1 through 3 provide insight into the extent to which ability estimates based on dichotomous and polytomous scoring converge. These figures plot two $\hat{\theta}$ for each examinee. The first estimate is based on traditional scoring, which assumes local item dependence holds for all items. The second estimate is based on polytomous scoring of the passages within each test section. For all test sections, the two estimates are very highly correlated (the lowest of the six

correlations was 0.975). However, dispersion is clearly seen in these figures. Recall that polytomous scoring allows test developers to treat a set of interrelated items as a single testlet, restructuring the test to minimize item dependencies. The only differences are (a) grouping of items into testlets in one of the scoring methods, and (b) the use of scoring weights reflecting the discriminating power of either items or testlets. Clearly, the two scoring methods produce different results.

Figure 1 plots the two estimates for Form 1 of the Biological Sciences section. Although the estimates are highly correlated (.99 for each form) and visually seem to produce similar results, some disparities are evident. For some examinees, even those in the middle of the ability distribution where measurement errors tend to be lower, ability estimation differences of almost one standard deviation are present. That such differences can be found across IRT ability levels is cause for concern. A difference of one standard deviation due to the choice of scoring method will have a highly significant impact on percentile rank, performance classification, and other important uses of the scores.

Figure 2 presents the scatter plot for Form 1 of the Physical Sciences section. The correlation for the two ability estimates across scoring methods on this test section is .987, marginally lower than on the Biological Sciences test section. Again, some of the ability estimates obtained from the two scoring methods differ by more than one standard deviation. Although the largest differences occur for examinees of low ability, difference of one standard deviation or more are noted throughout the plot.

For the Verbal Reasoning test section, where greater levels of LID were detected, the disparities between the two ability estimates are greater. Figure 3 presents the ability estimate scatter plot for Form 1, where the two estimates correlated .975. The scatter plot exhibits a

Figure 2. Plots of Ability Estimates: Dichotomous and Polytomous Scoring
Physical Sciences Form 1 (Correlation: 0.987)

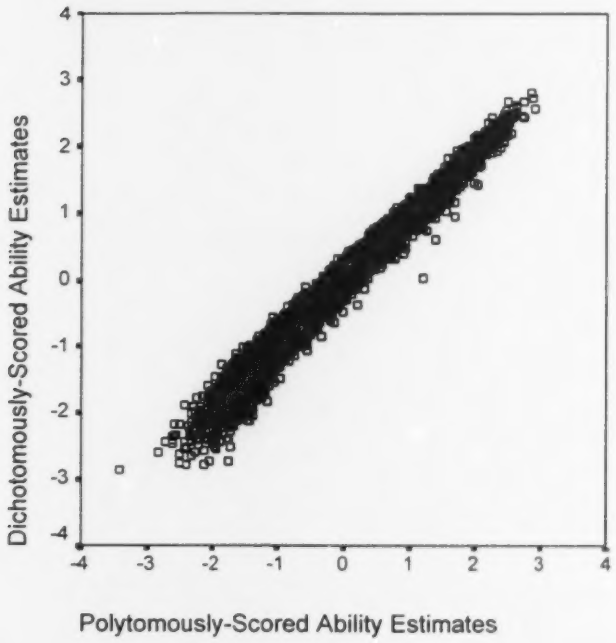
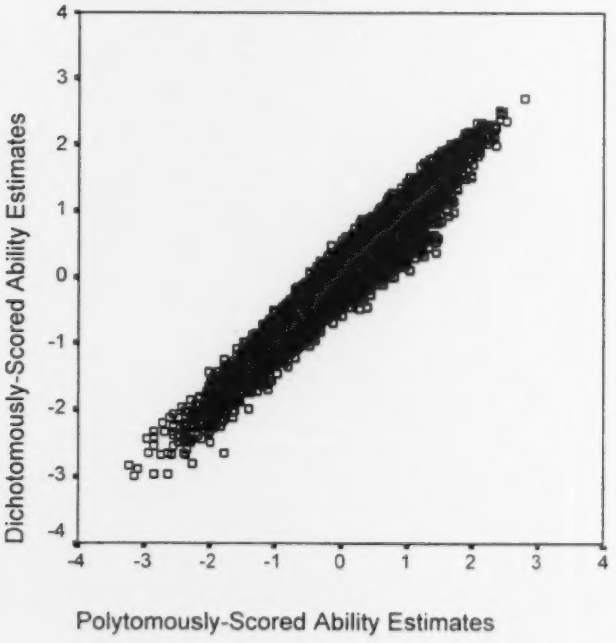


Figure 3. Plot of Ability Estimates: Dichotomous and Polytomous Scoring
Verbal Reasoning Form 1 (Correlation: .975)



noticeable "bulge" in the middle-to-upper area of the plot. Some ability estimates (albeit a small proportion) differ by nearly two standard deviations. The implication of such differences is that LID among items within passages cause problems for ability estimation. The end result of such differences is that for a sizable number of examinees, the choice of scoring method seriously impacts the score they will receive.

Discussion

With the use of assorted item formats (including sets of items linked to a common passage) that provide examinees with opportunities to showcase diverse skills, a number of novel scoring formats are also being developed. As these different item and scoring formats are incorporated into established tests, research in detecting LID should also be completed as a vital component of test reliability and validity.

Several interesting empirical findings relating to LID emerged in this study. A number of practical and easy-to-implement strategies for detecting dependencies already exist, although interpretation of these statistics remains somewhat problematic. Comparing reliability estimates across testlet and non-testlet scoring of context-dependent item sets is one way of determining if LID is present. However, the Q_3 statistic is more useful for identifying specific pairs of items that are locally dependent. As noted earlier, these statistics are descriptive, not statistical. Their magnitude often appears quite small (indeed, even the largest values cited in the literature are around .10), introducing added difficulty in interpreting their practical meaning.

With respect to the MCAT sections analyzed here, the results suggest some dependencies in the dichotomously scored item data. Two factors could underlie this dependence: speededness and context-dependence (related to passage-structure). A largely contextual explanation is called for on two of the test sections: Biological and Physical Sciences. Passages of the Problem

Solving type (and to a lesser extent, the Persuasive type) tended to exhibit more LID than other passage types for the Biological Sciences test section, while on Physical Sciences the Persuasive passages had the highest values. Many item pairs within these passages had noticeably larger Q_3 values. On the Verbal Reasoning test section, the results are more equivocal. The Q_3 statistics were comparatively higher across all testlets on this test section, and even slightly larger still toward the end of the test section. This indicates a combination of speededness and passage-related dependence.

Results from the ability estimation analyses indicated the dependencies observed on the three test sections may have practical consequences for ability estimation. As illustrated in Figures 1 through 3, choice of scoring method can have a significant impact on ability estimation for at least some candidates. If this were not the case, the bivariate plots would be fit perfectly by a straight line. This impact was especially noticeable on the Verbal Reasoning section, where item dependencies were most evident.

Methods for addressing the practical effects of LID are worthy of more investigation, for on any test where passages and item sets are used, associated item dependencies can seriously impact both the statistics we work with in test design and the scores that are ultimately reported to examinees. One area of future research is in designing field-tests of different versions of context-dependent item sets that could shed light on LID and how it should be modeled. Currently, many testing organizations field-test different sets of items associated with a common passage. The items that survive the field test may not have all appeared on the same field-test form. Thus, more work needs to be done to investigate whether certain combinations or orderings of items within an item set may alleviate LID. The use of an IRT model that is not based upon the restrictive assumption of local independence (Jannarone, 1991) is another

possible direction. Another potential direction for future research is investigation into the effect of scoring on predictive validity. For example, it may be interesting to study whether differences between testlet and discrete scoring of context-dependent item sets lead to differences in the predictive utility of test scores.

References

Anastasi, A. (1961). Psychological testing (2nd ed.). New York: Macmillan.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis. Cambridge, MA: MIT Press.

Chen, W. (1998). IRTNEW: A computer program for the detection of local item dependence. Chapel Hill, N.C.: L. L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22 (3), 265-289.

Fennessy, L. M. (1995). The impact of local dependencies on various IRT outcomes. Unpublished doctoral dissertation, University of Massachusetts at Amherst. [Dissertation Abstracts International, 56-03A, p.899.]

Ferrara, S., Huynh, H., & Bagli, H. (1997). Contextual characteristics of locally dependent open-ended item clusters on a large-scale performance assessment. Applied Measurement in Education, 12, 123-144.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large-scale hands-on science performance assessment. Journal of Educational Measurement, 36, 119-140.

Green, B. F., Bock, R. D., Humphreys, L. D., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.

Guilford, J. P. (1936). Psychometric methods (1st ed.). New York: McGraw-Hill.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, NJ: Sage.

Huynh, H., & Ferrara, S. (1994). A comparison of equal percentile and partial credit equatings for performance-based assessments composed of free-response items. Journal of Educational Measurement, 31, 125-141.

Jannarone, R. J. (1991). Conjunctive measurement theory: Cognitive research prospects. In M. Wilson (Ed.), Objective measurement: Theory into practice. Norwood, NJ: Ablex.

Kelley, T. L. (1924). Note on the reliability of a test: A reply to Dr. Crumm's criticism. The Journal of Educational Psychology, 15, 193-204.

Kingston, N. M., & Dorans, N. J. (1982). The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test (ETS Research Report 82-12). Princeton, NJ: Educational Testing Service.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. Applied Measurement in Education, 12, 237-255.

Lord, F. M., & Novick, M. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. Applied Psychological Measurement, 18(13), 245-56.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika, Monograph Supplement, No. 17.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28(3), 237-247.

Tang, K. L., & Eignor, D. R. (1997). Concurrent calibration of dichotomously and polytomously scored TOEFL items using IRT models (TOEFL Technical Report). Princeton, NJ: Educational Testing Service.

Thissen, D. (1991). MULTILOG 6.3 [Computer program]. Mooresville, IN: Scientific Software.

Thissen, D., Billeaud, K., McLeod, L., & Nelson, L. (1997, August). A brief introduction to item response theory for items scored in more than two categories. Paper presented at the National Assessment Governing Board Achievement Levels Workshop, Boulder, CO.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. Journal of Educational Measurement, 26, 247-260.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), Educational measurement (pp. 560-620). Washington, D.C.: American Council on Education.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. Applied Measurement in Education, 8 2), 157-186.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized-adaptive testing: A case for testlets. Journal of Educational Measurement, 24(3), 185-201.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? Educational Measurement: Issues and Practice, 15, 22-29.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30(3), 187-213.

AAMC ASSOCIATION OF AMERICAN MEDICAL COLLEGES
2450 N Street, NW, Washington, DC 20037-1127
Phone 202-828-0400 Fax 202-828-1125
www.aamc.org